Gaussian Mixture Models for Temporal Depth Fusion

Cevahir Cigla Roland Brockers Larry Matthies

Cevahir.Cigla@jpl.nasa.gov Roland.Brockers@jpl.nasa.gov lhm@jpl.nasa.gov

Jet Propulsion Laboratory, California Institute of Technology

Abstract

Sensing the 3D environment of a moving robot is essential for collision avoidance. Most 3D sensors produce dense depth maps, which are subject to imperfections due to various environmental factors. Temporal fusion of depth maps is crucial to overcome those. Temporal fusion is traditionally done in 3D space with voxel data structures, but it can be approached by temporal fusion in image space, with potential benefits in reduced memory and computational cost for applications like reactive collision avoidance for micro air vehicles. In this paper, we present an efficient Gaussian Mixture Models based depth map fusion approach, introducing an online update scheme for dense representations. The environment is modeled from an ego-centric point of view, where each pixel is represented by a mixture of Gaussian inverse-depth models. Consecutive frames are related to each other by transformations obtained from visual odometry. This approach achieves better accuracy than alternative image space depth map fusion techniques at lower computational cost.

1. Introduction

Depth perception is fundamental to most approaches to obstacle detection for robotic vehicles, and reliable obstacle detection is particularly challenging for small micro air vehicles, which are the main application focus here. Significant research has been devoted to dense depth perception with stereo matching [1]-[3] and active sensors, such as Microsoft Kinect, Intel RealSense, and Time-of-Flight cameras. Despite this, depth map errors are still frequent, generally due to the presence of non-Lambertian surfaces, textureless regions, changes in lighting that have uneven effects on the scene, and inherent range limitations of active depth sensors. Obstacle detection errors -- false alarms and missed detections -- are inevitable if detection is only done with instantaneous frames of depth data.

Naturally, such errors can be reduced by temporal fusion. In the robotics literature, temporal fusion in 3D space with occupancy grid or voxel data structures has been a standard approach [4]-[6]. However, temporal fusion can also be done in image space (Figure 1). This has potential to reduce obstacle detection error rates at lower computational cost, particularly for reactive navigation in cluttered environments. With inverse range as the depth parameterization, image space temporal fusion also avoids problems with defining appropriate 3D map cell sizes when the uncertainty of depth measurements is a strong function of the true depth, as is the case for many sensors. Image space fusion could also be a useful front end to quickly filter inconsistent depth measurements before creating a 3D world model.

The research on depth enhancement has mostly focused on spatial enhancement such as joint depth-color filtering [7]-[10] and up-scaling [12]-[16]; while temporal enhancement has been given much less attention. The large literature of simultaneous localization and mapping (SLAM) can be considered a way of fusing temporal data in order to generate a representation of an environment. However, the sparse representation of these techniques is not appropriate for path planning and collision avoidance that require denser representation [17]-[18]. The multi-view 3D extraction techniques can be adapted to temporal domain by assigning consecutive frames as multiple observations of a scene. This approach can provide temporal consistency while demanding high computational power due to use of multiple 3d warping.

In this paper, we propose an efficient depth data fusion technique to provide temporally consistent models for path planning and collision avoidance for ground vehicles or micro air vehicles. Our solution is inspired by a background modeling framework for surveillance image change detection, where each pixel is represented as a mixture of Gaussian distributions. This compact depth map representation is propagated between frames by forward warping, using platform ego motion estimates, and is updated at each time step using newly observed depth maps. Assuming the rigid scene with low level of moving objects, depth maps can be provided by an active sensor, stereo matching or structure from motion.

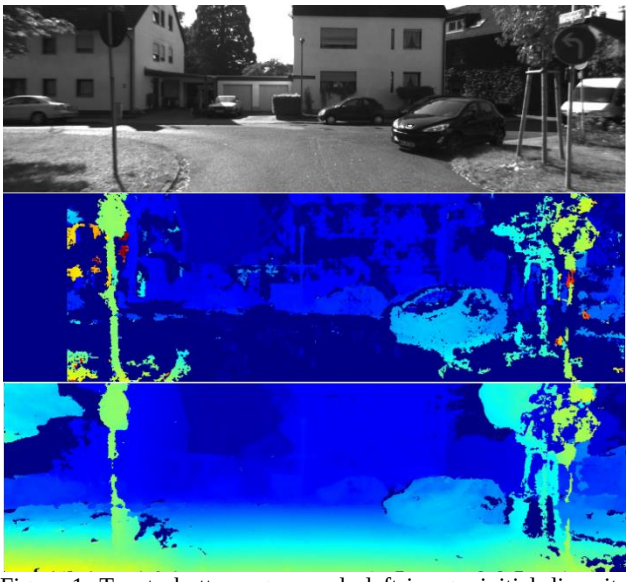

Figure 1: Top to bottom: gray-scale left image, initial disparity map via Semi Global Matching, temporally fused disparity map. Temporal fusion compensates flickers, un-reliable disparity estimates and empty pixels for denser representation of the surrounding.

The remainder of the paper is organized as follows. The next section summarizes prior work related to temporal fusion. Section 3 presents the details of the proposed approach, followed by experimental results and comparison to other methods in Section 4. Finally, we discuss conclusions and potential future directions in Section 5.

2. Related Work

Temporal fusion of depth data can be classified into three categories. The first group integrates temporal consistency in the cost function during the extraction of 3D. [19] exploits Markov Random Fields constructed on multiple consecutive frames. The independently extracted depth maps are merged through bundle optimization resulting in high computational complexity. In [20] monocular dense reconstruction is proposed by describing each pixel as a parametric model to extract depth maps from a multi-view stereo point of view. The approach presented in [20] differs from traditional multi-view stereo techniques by introducing online and sequentially updated depth maps. In [21]-[23], local edge-aware filters over temporally aggregated cost functions are utilized to determine the depth maps. The SLAM literature, while using sparse representation, exploits online depth updates especially in the key frames. In [24][25], the sparse depth measurements are modeled as a weighted sum of Gaussian and uniform distributions, and the depth search is performed along a restricted region in the epipolar line. On the other hand, in [26] a simple Gaussian model is utilized to model depth measurements and the depth search is limited within the standard deviation of the prior hypothesis. The depth update is achieved by multiplications of two distributions as in the Kalman filter update step. This approach is extended to large scale direct SLAM with the addition of stereo cameras in [27], where stereo matching is exploited to adjust the monocular scale and increase the number of reliable points. In SLAM techniques, occluded pixels are eliminated from the model according to the variance of the depth values.

The second group relies on the utilization of 3D models, such as voxels or surfaces, to fuse depth data. KinectFusion [28] gets the depth maps from Kinect camera with active sensors and these maps are merged through signed distance functions to efficiently represent the 3D surfaces. In RGB-D Fusion [29], the depth and color data captured from RGB-D sensors are merged to increase accuracy of the 3D models. These approaches exploit high power GPUs to meet high precision and real-time requirements. They are generally applicable to indoor 3D model reconstruction that limits the scope of path planning and collision avoidance. In [31], the raw depth values gathered from active sensors are improved via median filter among nearby frames in a time window**.** [30] models depth on rays in occupancy map as a mixture model.

The final group involves techniques that approach the temporal fusion problem in a post-processing or filtering framework. [32] proposes a visibility based approach to fuse multiple depth maps into a single depth map. This method requires multiple 3D warping and depth ordering steps for each frame that increases the memory and computation requirement. The resulting depth maps still include noise, since visibility is constrained for each frame independently without any global regularization. In [33], depth maps are integrated into a volumetric occupancy grid. Two level height maps are exploited to constrain the motion of a robot in an indoor environment. The regularization is achieved through anisotropic total variation with reduced dimension due to indoor constraints. In [34], depth estimates of consecutive frames are merged by a probabilistically motivated 3D filtering approach. Each frame receives multiple depth candidates from the preceding frames and the depth assignment is achieved by maximization over the local histogram of mean-shift filtered depth values. This merging step is followed by photometric edge-aware filtering and mesh generation to fill the holes in the depth maps. [35] utilizes a median filter over consecutive frames to smooth out the noisy measurements then averages the depth values according to the motion estimation between color images and inter-frame differences. In [16], optical flow and patch similarity measures are exploited to up-scale low resolution ToF cameras w.r.t. high resolution color images and provide temporal consistency. [36] projects multiple depth hypotheses to a reference view and estimates probability density function of depth measurements via projection uncertainties. The depth candidate with highest probability

is assigned to the corresponding pixel. Recently, [37] proposes a novel total generalized variation technique to fuse depth maps from multiple frames. The optimization is executed on a 2.5 D surface obtained by back-projecting the depth maps.

It is important to note that using multiple 3D warpings (back-projection and projection) or optical flow are the two alternatives for data registration for the techniques that consider fusion as a filtering framework. This is a limiting factor in terms of memory and computational complexity for onboard processing. Even the median filter, a common approach to remove outliers, requires high computation. In addition, multi-view techniques suffer from holes created during forward mapping as the motion between frames increases.

In order to provide temporally consistent disparity maps and denser representation, we propose a sequential depth map filtering approach where each pixel is considered as a mixture of Gaussian models. We consider the problem from an egocentric point of view by only considering the current field of view and ignoring the previously visited out-of-view regions. This compact representation yields an efficient solution to address the trade-off between computational complexity and accuracy. GMMs are projected onto the most recent frame with respect to pose estimates gathered from a SLAM framework. Hence, only the pose change between the recent two frames is exploited, which reduces the required number of 3D warpings tremendously. The Gaussian models are updated efficiently with the current depth map observation. This approach unites and extends the efficiency of sparse depth model updates in the SLAM literature with dense representation of multi-view stereo. The use of Gaussian mixtures enables modeling partially occluded pixels due to ego-motion or independently moving objects.

3. GMM based Temporal Fusion

Use of Gaussian Mixture Models (GMM) is a common technique to perform background/foreground segmentation for detecting moving objects in surveillance video [38]. This approach combines sequential observations of a pixel (intensity) in a compact representation. The same idea can be extended to represent the environment on a moving platform; we address this formulation here. This is closely related, but not identical to, formulations that would result from a strict recursive state estimation derivation. We show that the GMM-inspired formulation is an advance over previous work, and expect to continue to examine variations in ongoing work.

3D sensors produce depth maps in the image domain, and are subject to errors and missing data due to many causes. Even where depth estimates are approximately correct, for many sensors the error in estimated 3D coordinates is a strong function of the true range; for example, this error is quadratic in range for triangulation-based sensors and nonlinear in range for phase-based time-of-flight active optical range sensors. This nonlinear error characteristic complicates the definition and maintenance of 3D grid-based world models. Similarly, most 3D sensors have angular instantaneous fields of view (IFOV), e.g. the projected cone imaged by one pixel, which also leads to sampling issues with 3D grid-based world models. Representing uncertainty in inverse depth in image space avoids these problems. However, gross errors from several sources can lead to ambiguous depth estimation given time sequences of observations; the GMM formulation offers a compact, efficient approach to overcome this ambiguity.

3.1. Notation

Let $\vec{x} = (u, v, d)$ be the triplet defining pixel position *(u,v)*, and the disparity value, *d*, in the image domain. We assume that, at a time instant *t*, x has a mixture of *K* Gaussian distributions as follows:

$$P(\vec{x}_t|X_T) = \sum_{m=1}^{K} w_m N(\vec{x}; \vec{\mu}_m, \vec{\sigma}_m) \qquad (1)$$

where $\vec{\mu}$'s are the mean and $\vec{\sigma}$'s are the variance estimates of the $\vec{x}$ triplet and X_T is the set of observations within time frame of T from the image sequence. In typical GMM applications, each mode has a weighting factor that affects the state of the pixel. In the depth integration version of this model, we exploit an occurrence counter on each mode and decide the current state w.r.t. occurrence and variance estimates. The variances of *u* and *v* are ignored, for the sake of efficiency, since these positions are only utilized to map points to the following frames without suffering quantization noise. Hence the variance of positions does not have a direct effect on the disparity values. Therefore, GMM is modified as follows:

$$P(\vec{x}_t|X_T) = \sum_{m=1}^{K} W(O_m, \sigma_m) N(\vec{x}_t; \vec{\mu}_m, \sigma_m) \quad (2)$$

where O_m corresponds to the number of frames that the corresponding mode *m* is observed and *W* is a weighting function that defines the contribution of the corresponding mode w.r.t. occurrence and variance. In this study *W* is chosen to be an impulse function centered at the mode with lowest variance and sufficiently high occurrence count. This choice provides crisp disparity refinement and handles the mixing of background and foreground hypotheses.

Visual SLAM pose estimates between consecutive frames provide the mapping of a triplet in frame *t-1*, to the following frame, *t*, as;

$$\vec{x}_t^h = {}_{t-1}^{t}\Theta(\vec{x}_{t-1})\, \vec{x}_{t-1} \qquad (3)$$

where ${}_{t-1}^{t}\Theta(\vec{x}_{t-1})$ is the *4x4* transformation matrix that

maps $\vec{x}_{t-1}$ to the following frame, and $\vec{x}_t^h$ is the hypothesized model. The mapping between two consecutive frames requires an inverse projection from the image coordinates to 3D, then a transformation based on the camera motion and a re-projection.

3.2. GM Modeling

GMM based temporal fusion involves initialization, forward mapping GMM update, and disparity assignment steps. Initialization create a single mode for each pixel *(x,y)* as follows:

$$N(\vec{x};\vec{\mu}_0,\sigma_0):\begin{cases}\vec{\mu}_0=(x,y,d)\\ \sigma_0=\sigma_{init}\\ O_0=1\end{cases} \quad (4)$$

In (4), σ_{init} is set to a high value (i.e., 6), and d is the observed disparity map at the initial frame. The initial high standard deviation indicates that the disparity value that is observed for first time is not trusted. The forward mapping step transfers models from the previous frame to the current frame and sets the valid disparity hypotheses for each pixel. Then, the update step fuses the temporally aggregated models with observation from the current disparity map. Finally, the assignment step outputs a single disparity estimate for each pixel by assessing the mixture distributions at each pixel.

3.2.1 Forward Mapping

At each time step, GMMs from the previous time step are mapped to the current time according to (3). This forward mapping is provided for all of the models
of a pixel. Therefore, the maximum number of 3D warpings is limited by the pre-defined number of models in the mixture, *K*. Forward mapping may introduce some holes due to quantization and occlusions as a result of the motion of the vehicle. The size of the holes is a function of the vehicle motion between frames; for large motions, some fundamental tools of forward mapping, such as dilation-erosion and Z-buffering, are not applicable since they are utilized when the source pixels have one disparity at a time. In this case, each pixel has a different number of depth models, which results in multiple depth models in the target frame. Thus, there is not a specific disparity map for applying any post-processing. Moreover, GMM depth models are considered to store partially occluded pixels along then temporal axis, hence exploiting a Z-buffer is not attractive, because it eliminates the occluded disparity candidates.

This problem is handled by allowing the mapped triplets to influence neighboring pixels in the target frame. Since each pixel gets contributions from the neighboring pixels, this increases the number of depth hypothesis. The number of possible hypotheses is limited by the predefined number of GMMs, *K*. Hence, a reduction step is used that averages triplets whose disparity hypothesis are closer than a threshold, i.e. $\Delta d=3$. The averaging is performed on all parameters of GMMs as follows:

$$N(\vec{x}^h;\vec{\mu}_m,\sigma_m)=\frac{1}{P}\sum_{s\in S}N(\vec{x}_s^h;\vec{\mu}_s,\sigma_s) \quad (5)$$

where $S=\{\vec{x}_s^h:|\vec{\mu}_m-\vec{\mu}_s|<\Delta d\}$ is the set of neighbor hypotheses and $p=|S|$ is the size of the set. The reduction is finalized by picking the best *K* models according to their standard deviations. This approach fills quantization holes, but it may grow object boundaries. This growing is handled by the rejection step during update of GMMs, which is explained in the following sub-section.

3.2.2 GMM Update

As a new frame is observed *(x,y,d)*, a comparison is conducted between the current disparity map, $\vec{x}(d)$, and the mapped GMMs from the previous frame as:

$$M=\underset{m\in[1,K_x]}{\operatorname{argmax}}|d-\vec{\mu}_m(d)| \quad (6)$$

In (6), the mode with closest disparity model is determined among the K_x prior models of the corresponding triplet. If the best match has disparity distance below a specified threshold, T_d, then it is considered to be a proper fit. In that case the update of GMMs, a common way for background update [38], is achieved as follows:

$$\begin{aligned}\sigma_M^2&=\alpha\sigma_M^2+(1-\alpha)|d-\vec{\mu}_M(d)|^2\\ \vec{\mu}_M&=\alpha\vec{\mu}_M+(1-\alpha)\vec{x}\\ O_M&=O_M+1 \qquad m\in[1,K_x]\\ \sigma_m^2&=\sigma_m^2+V_0\\ O_m&=O_m-1\end{aligned} \quad (7)$$

where the matched mode, *M*, is updated by the current observation. The remaining modes are penalized by V_0 (=0.5) since they do not have any observations. In addition, the occurrence counter is incremented for the matched mode, M, and decremented for the mismatched modes. The update rate, α, is fixed at a value that balances rapid convergence with smoothing over many frames. Experiments show that this update process improves performance over prior work at lower computational cost; future work will examine alternate probabilistic foundations of the update formulation.

If there is no fit, all GMMs of the corresponding pixel are penalized and a new mode is included according to (4). If the number of modes is at the limit, the weakest mode (w.r.t. disparity variance) is replaced with the current observation. There may be no observation coming from the current disparity map; in that case, the models are not

updated while the occurrence count is decreased as a forgetting factor. In order to adapt to temporal changes and preserve efficiency, modes with high occurrence counts but large disparity variances are rejected. These correspond to unreliable modes, since the variances have not decreased despite high occurrence.

3.2.3 Disparity Assignment

For motion planning, each pixel is assigned a final disparity estimate according to the occurrence count and the variance of the GMMs. To assign a valid disparity value, the mode that fits the most recently observed disparity must have an occurrence count larger than a threshold (i.e. $O_m > 3$). This rejects temporal flickers among consecutive frames. Also, the variance estimate of the model should also be below a threshold, which enforces the assignment to be reliable ($\sigma_m < 0.25\sigma_{init}$). The same conditions are valid for the empty pixels or when there is no match with the prior GMMs in the current disparity map. In this case, the best model having least disparity variance is assigned to the corresponding pixel as long as the occurrence and variance satisfy the conditions.

3.3. Computational Analysis

Before describing experimental results, we give a brief analysis of the computational complexity of the proposed fusion technique. Common multi-view techniques where fusion is considered as a filtering problem are considered as a baseline for comparison. Neglecting the post-processing steps and additional optimizations, the comparison is based on the required number of forward mappings and the memory requirement to hold multiple hypotheses. This gives a general idea of the complexity without getting into details of additional processes. In the multi-view approach, the number of 3D warpings is at least equal to the width of the time window, given as *T*, and the memory requirement is *TxWxH* to store all possible contributions from the previous frames. On the other hand, the proposed approach requires *K* 3D mappings and *5K* of image memory (three for triplet means (*u*,*v*,*d*), one for occurrence count and one for disparity variance). Single 3D mapping as given in (3) involves two projections in addition to one transformation in 3D coordinates. In the stereo camera setup, the projections are simplified by basic arithmetic operations over the camera calibration parameters.

In general, 10 to 20 frames are utilized during multi-view depth fusion [34][36][37], while for GMM based fusion one to three GMMs are sufficient to provide a compact representation of the previous depth maps. Hence, there is an obvious decrease in the number of 3D forward mappings, which is a time-consuming step especially for on board processing. On the other hand, the memory requirement remains on the same scale.

4. Experimental Results

To evaluate the performance of the proposed approach, we utilize the well-known KITTI stereo benchmark [3]. This provides an excellent framework to evaluate stereo matching and multi-view extensions, since it contains sets of 20 consecutive test frames, with ground truth at 11th frame; hence, 10 frames are utilized as the time window for temporal fusion.

In the first set of experiments, comparative results are given with state-of-the-art techniques in terms of computation time and depth map accuracy. We fixed the number of modes at *K=3* and neighbor set as *S:3x3* for the proposed approach, which provides a good tradeoff between computation and accuracy. In the second set of experiments, we analyze the performance of GMM based fusion with respect to number of modes in the mixture and different visual odometry poses gathered from three different approaches [39]-[41].

Throughout the experiments, two different stereo matching algorithms (Semi-Global Matching (SGM) [42] and Efficient Large Scale Stereo Matching (ELAS) [43]) are exploited to observe the enhancement after temporal fusion. Both of the matching techniques yield sub-pixel estimates, so the disparity values after temporal fusion also have sub-pixel accuracy. The parameters for the stereo matching algorithms are set according to the parameter set given KITTI evaluation benchmark. The evaluation is based on the mean disparity error and the percentage of erroneous pixels with disparity error larger than a threshold, i.e., $\Delta d > 3$.

4.1. Comparison with State of the art

We selected TGV [37], PFuse [36], DSM [34] and the common median filter as the techniques to compare. In all these techniques, the problem is considered in a filtering framework, as in our approach, where fusion is conducted with estimated disparity maps and camera poses estimated by VO. TGV is a complex optimization framework with high ranking on the KITTI test benchmark, thus it is considered as state-of-the-art in terms of accuracy. Sharing the same experimental setup, we quote published results of TGV and report results of PFuse, DSM and median filter obtained by our implementation. Post-processing steps of these algorithms are not exploited in order to evaluate the fusion stage only. The tests are conducted on the KITTI stereo 2012 training set, including 194 different sequences with average resolution of *1250x350*. This set has mostly static scenes, compared to the 2015 release of the benchmark that has independently moving objects at each frame. Thus, the KITTI 2012 set provides a more focused evaluation of temporal fusion.

In this set-up, visual odometry pose estimates are obtained via [39]. The performances of the temporal fusion algorithms over the disparity maps obtained by SGM and

ELAS are given in Table 1 and Table 2, respectively. In both cases, the best average disparity error (D-avg) and best outlier percentage (Out-3%) are achieved by the proposed GMM-based fusion approach with *K=3*. Performance of temporal fusion is generally less with ELAS than with SGM, because SGM results start out worse. For empty pixels with no disparity assignment, background filling is performed before the evaluation. The proposed technique gives better error percentages than the other methods. PFuse and DSM perform poorly compared to even median filtering due to fast motion of the vehicle. DSM is designed for images captured with down looking cameras on airborne platforms and PFuse for parking assistance; thus, both require small disparity changes between consecutive frames.

Apart from the accuracy, density of the proposed approach is lower due to hard constraints (introduced in section 3.2.3) to reduce the temporal flickering effect and increase reliability. Completeness of the fused disparity maps could increase by decreasing the thresholds; however, in that case temporal consistency would slightly decrease. We set the thresholds such that a minimum number of outliers is observed. In Figure 2 and Figure 3, disparity maps of some selected scenes are illustrated, which support the results presented in Table 1 and Table 2. We cannot show results of TGV [37], since they are not available. However, we can assume they are visually similar to the proposed approach given the similar the numerical results. Especially for cases where standard stereo matching fails due to change of lighting and reflection, temporal fusion handles this and propagates the previous reliable disparity hypotheses to the unreliable pixels. Outlier spikes in the disparity maps are minimal in the proposed approach, whereas we observe more spikes in PFuse, DSM, and Median, especially in Figure 3. In general, the simpler and lower cost the base stereo algorithm, the more benefit we expect will be obtained from temporal fusion; thus, inexpensive local block matching stereo algorithms should benefit even more.

The proposed approach preserves crisp objects boundaries such as the traffic sign and the pole; on the other hand, objects are enlarged by the other techniques. On the traffic sign in Figure 2, the background is mixed for the remaining techniques, while the proposed approach preserves the valid disparity estimate. This is achieved by the accumulation of recent observations close to camera through neighbor influence that removes holes in the disparity maps.

The average execution times are given in Table 3. The timing for TGV-1 is copied from the related paper, which used GPU implementation for a significant speed-up. The remaining techniques were tested on a 3.4 GHz i7-3770 CPU. For the rest of the algorithms, the same forward warping tool is exploited to be fair with no additional optimizations. A 3x3 closing operation is implemented for

Table 1: The average disparity error and out-3 percentage performances of the temporal fusion techniques over SGM

Err > 3	**D-avg**	**Out-3 %**	**Density%**
SGM [42]	2.9	13.1	76
TGV [37]	2.0	8.6	**100**
PFuse[36]	2.5	11.5	93
DSM [34]	2.6	12.0	97
Median	2.1	9.1	99
Proposed	**1.8**	**7.9**	94

Table 2: The average disparity error and out-3 percentage performances of the temporal fusion techniques over ELAS

Err > 3	**D-avg**	**Out-3 %**	**Density %**
ELAS [43]	1.7	9.8	76
TGV [37]	1.4	7.3	**100**
PFuse [36]	1.8	8.9	92
DSM [34]	1.9	9.5	99
Median	1.5	7.2	99
Proposed	**1.3**	**7.1**	92

PFuse, DSM and Median Filter to fill the holes to an extent in order to increase their performance. The timings in Table 3 can be further improved with additional optimization for onboard processing. In Table 3, efficiency of the proposed approach is clear and is a result of compact representation, decreased number of 3D warping steps and simple update steps. The processing time for PFuse and DSM are high; PFuse exploits additional 3D warping to test different disparity hypotheses on multiple images, while DSM uses an iterative mean-shift approach that is time consuming. The accuracy of temporal fusion improves as longer time windows are exploited. In this case, the proposed approach does not need additional computation, due to the online frame-at-a-time update, while the rest of the multi-view approaches require further computation.

In order to understand the effect of temporal fusion in detail, the percentages of pixels with different disparity errors are illustrated in Figure 4. In these plots, each color indicates the contribution of the pixels of an error region to the average error. For example, after proposed temporal fusion over SGM disparity maps, the pixels with $2 > \Delta d > 1$ (indicated by blue color) have the contribution of almost 60% to the average error *1.8*. Temporal fusion specifically decreases the number of pixels with high disparity errors. In the meantime, these pixels are shifted to lower error bands as observed by the enlarged percentage of pixels with $2 > \Delta d > 1$. The refinement is more visible if the initial disparity maps have higher error rates. The error sensitivity may change depending on the application, so providing a complete error distribution yields much clear understanding of the effects of temporal fusion.

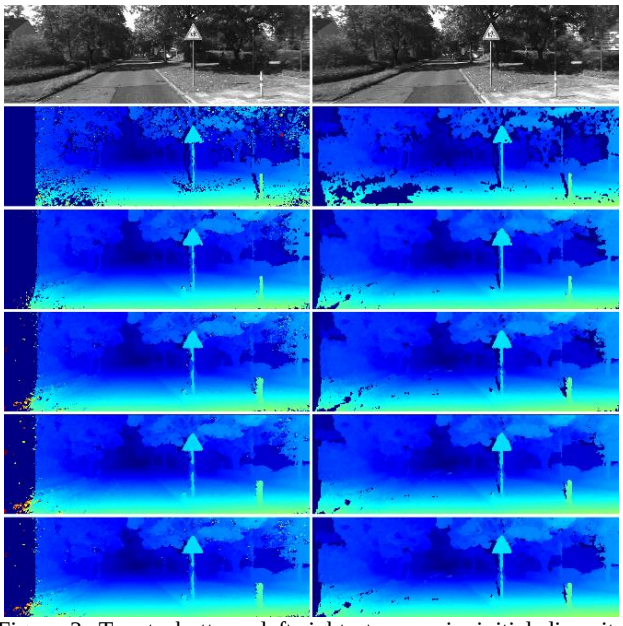

Figure 2: Top to bottom: left-right stereo pair, initial disparity maps (SGM left, ELAS right), proposed approach, PFuse [36], DSM [34]and Median Filter.

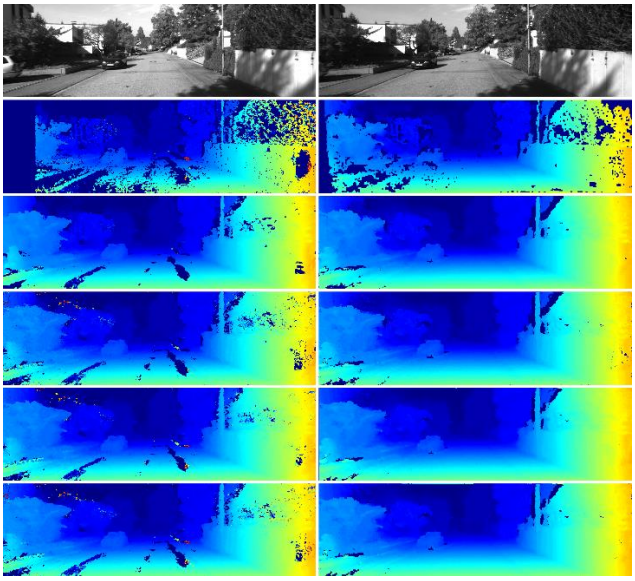

Figure 3: Top to bottom: left-right stereo pair, initial disparity maps (SGM left, ELAS right), proposed approach, PFuse [36], DSM [34] and Median Filter.

Table 3: The execution times of the algorithms

	Time (sec)	**Platform**
TGV [37]	70	GPU
PFuse [36]	23	CPU
DSM [34]	25	CPU
Median	1.3	CPU
Proposed	0.4	CPU

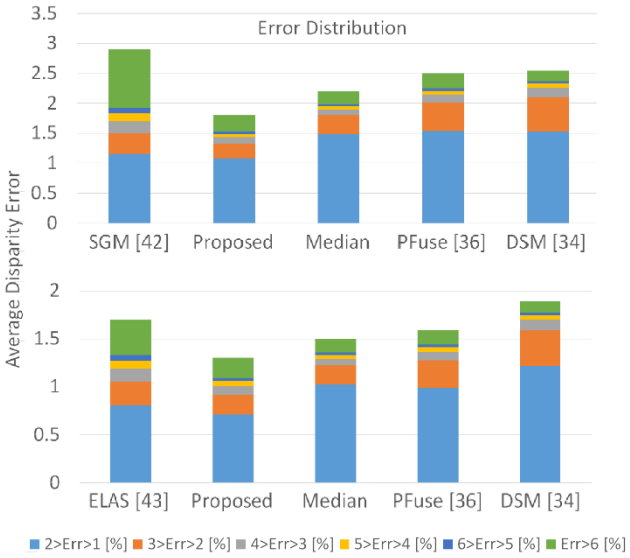

Figure 4: Distribution of errors according to different bounds.

4.2. Effects of parameters and VO

The most critical parameter for GMM based fusion is the number of GMMs, since that affects the model complexity and computation time. We extracted distributions of the number of GMMs for two long stereo sequences with 10000 frames from the KITTI odometry dataset [3], as well as the dataset used for stereo evaluation (2134 frames). On the average, the mode distributions are given in Table 4. The distribution of models is related to the complexity of the dataset. The odometry sequence involves more moving objects compared to stereo benchmark sequences, so the percentage of side modes is higher than the stereo benchmark set. Since the first three modes cover 90% of the distribution for the stereo benchmark, that is a good choice for algorithm parameter.

The distribution of erroneous pixels with different error thresholds is given in Table 5 for the proposed approach with three different limiting mode numbers over SGM [42] and ELAS [43] separated by slash respectively. This error representation yields a comprehensive understanding of the distribution of error. Performance is very similar for all three cases. One of the causes of this small performance difference is that the data set has static scenes.

In order to test the robustness of the proposed approach, the same experiments were conducted with three different VO algorithms [39]-[41] for 3-mode GMM fusion. Table 6 shows the error distributions as the percentage of pixels with quantized errors. The VO poses provided by [40] are improved in [39] which is a newer study. [41] has the worst VO pose estimates among the three, which are used in TGV [37]. According to the results in Table 6, the accuracy of VO poses affects the performance of temporal refinement, as expected. However, all VO poses result in same average disparity error and the differences for high error

percentages are almost insignificant. On the other hand, the difference for low error pixel rates is significant. These show that the proposed approach is robust against different visual odometry accuracy in terms of average disparity error as long as VO performs well enough to relate consecutive frames.

Table 4: Mode distribution over different stereo sequences

Mode Distribution	1 mode	2 mode	3 mode	4 mode	5 mode
Odometry	52%	23%	13%	7%	5%
Stereo	64%	18%	9%	5%	4%

Table 5: The percentages of error for different thresholds by use of different number of GMM modes over SGM/ELAS

%	[42]/[43]	1-mode	2-mode	3-mode
Out-1	27.1/25.8	30.3/24	30/23.9	29.9/23.8
Out-2	16.3/13.5	12.5/11	12/10.8	12.0/10.7
Out-3	13.1/9.8	8.3/7.4	7.9/7.1	7.9/7.1
Out-4	11.3/7.8	6.6/5.6	6.2/5.4	6.2/5.4
Out-5	10.0/6.5	5.7/4.6	5.3/4.5	5.3/4.7
Out-6	9.1/5.6	4.9/3.9	4.6/3.8	4.6/3.7
Davg	2.9/1.7	1.9/1.4	1.8/1.3	1.8/1.3

Table 6: The percentages of error for different thresholds for GMM based fusion, *K=3*, w.r.t three different VO poses

%	[42]/[43]	[39]	[40]	[41]
Out-1	27.1/25.8	29.9/24	30.2/24.3	30.7/24.7
Out-2	16.3/13.5	12.0/11	12.8/11.3	13.2/11.7
Out-3	13.1/9.8	7.9/7.1	8.5/7.6	8.7/7.9
Out-4	11.3/7.8	6.2/5.4	6.6/5.9	6.8/6.0
Out-5	10.0/6.5	5.3/4.4	5.6/4.8	5.7/4.9
Out-6	9.1/5.6	4.6/3.7	4.9/4.1	4.9/4.2
Davg	2.9/1.7	1.8/1.3	1.8/1.4	1.8/1.4

5. Conclusions

In this paper, we propose an efficient GMM inspired approach to fuse disparity maps temporally. Each pixel is represented by mixture of multiple models accumulated through previous observations. This compact representation is mapped to the following frame via the 3D transformation between camera poses. The models are utilized to refine the recent disparity observations and updated for the next frames. The online update approach fuses temporal data efficiently and does not require any time window. According to comprehensive experiments, the proposed approach is an efficient alternative for the state-of-the-art with far lower computational complexity and competitive accuracy. The proposed approach yields temporally consistent, flicker-free disparity maps with fewer errors and more complete representation, which are vital for collision avoidance. Use of multiple models may also enable the detection and segmentation of independently moving objects in complex environments, which remains as a future direction of this study.

6. Acknowledgement

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

[1] D. Scharstein and R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1/2/3):7-42, April-June 2002

[2] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling, High-resolution stereo datasets with subpixel-accurate ground truth. *German Conference on Pattern Recognition, September* 2014.

[3] A. Geiger, P. Lenz and R. Urtasun, Are we ready for Autonomous Driving? The KITTI Benchmark Suite. *Conference on Computer Vision and Pattern Recignition*, 2012

[4] A. Hornung,. K.M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, OctoMap: An Efficient Probabilistic 3D Mapping Framework Based on Octrees**.** *Autonomous Robots*, 2013

[5] D. Cole and P. Newman, Using Laser Range Data for 3D SLAM in Outdoor Environments. *IEEE International Conference on Robotics and Automation*, 2006

[6] I. Dryanovski, W. Morris and J. Xiao, Multi-volume Occupancy Grids: An Efficient Probabilistic 3D Mapping Model for Micro Aerial Vehicle. *International Conference on Intelligent Robotics and Systems*, 2010

[7] J. Dolson, J. Baek, C. Plagemann and S. Thrun, Upsampling Range Data in Dynamic Environments, *Conference on Computer Vision and Pattern Recognition*, 2010.

[8] Q. Yang, R. Yang, J. Davis, and D. Nister. Spatial-depth Super Resolution for Range Images. *Proc. Conference on Computer Vision and Pattern Recognition*, 2007.

[9] J. Park, H. Kim, Y. W. Tai, M. Brown, and I. Kweon, High Quality Depth Map Up-sampling for 3d-tof cameras. *International Conference on Computer Vision*, 2011.

[10] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof. Image Guided Depth Up-sampling using Anisotropic Total Generalized Variation. *International Conference on Computer Vision*, 2013.

[11] J. Kopf, M. F. Cohen, D. Lischinski and M. Uyttendaele, Joint Bilateral Up-sampling. SIGGRAPH 2007

[12] M.Y. Liu, O. Tuzel and Y. Taguchi, Joint Geodesic Upsampling of Depth Images. *Conference on Computer Vision and Pattern Recognition*, 2013

[13] N. Schneider, L. Schneider, P. Pinggera, U. Franek, M. Pollefeys and C. Stiller. Semantically Guided Depth Upsampling. *German Conference on Pattern Recognition*, 2016

[14] J. Lu, D. Forsyth, Sparse Depth Super Resolution, *International Conference on Computer Vision and Pattern Recognition*, 2015

[15] K. Matsuo, Y. Aoki, Depth Image Enhancement Using Local Tangent Plane Approximations. *International Conference on Computer Vision and Pattern Recognition*, 2015

[16] D. Min, J. Lu and M. N. Do, Depth Video Enhancement Based on Weighted Mode Filtering. *IEEE Transactions on Image Processing, 21(3), March* 2012

[17] K. Schmid, T. Tomic, F. Ruess, H. Hirschmuller and M. Suppa, Stereo Vision based indoor/outdoor Navigation for Flying Robots. *International Conference on Intelligent Robots and Systems*, 2013.

[18] L. Matthies, R. Brockers, Y. Kuwata and S. Weiss, Stereo Vision-based Obstacle Avoidance for Micro Air Vehicles Using Disparity Space. *IEEE International Conference on Robotics and Automation*, 2014.

[19] G. Zhang, J. Jia, T. T. Wong and H. Bao, Consistent Depth Maps Recovery from a Video Sequence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6), June 2009

[20] M. Pizzoli, C. Forster and D. Scaramuzza, REMODE: Probabilistic, Monocular Dense Reconstruction in Real Time. *IEEE International Conference on Robotics and Automation* 2014.

[21] C. Richardt, D. Orr, I. Davies, A. Criminisi and N. A. Dodgson, Real-time Spatiotemporal Stereo Matching Using the Dual-Cross-Bilateral Grid, *European conference on Computer vision*, 2010

[22] A. Hosni, C. Rhemann, M. Bleyer, M. Gelautz, Temporally Consistent Disparity and Optical Flow via Efficient Spatio-Temporal Filtering. *Pacific-Rim Synmposium on Image and Video Technology*, 2011.

[23] C.C.Pham, V. D. Hguyen and J. W. Jeon, Efficient Spatio-Temporal Local Stereo Matching Using Information Permeability Filtering. *International Conference on Image Processing*, 2012.

[24] C. Foster, M. Pizzoli and D. Scaramuzza, SVO: Fast Semi-Direct Monocular Visual Odometry. *IEEE International Conference on Robotics and Automation*, 2014

[25] G. Vogiatzis and C. Hernandez, Video-based, Real-Time Multi View Stereo. *In Image and Vision Computing*, 2011.

[26] J. Engel, J. Strum and D. Cremers, Semi-Dense Visual Odometry for a Monocular Camera. *IEEE International Conference on Computer Vision*, 2013

[27] J. Engel, J. Stueckler and D. Cremers, Large-Scale Direct SLAM with Stereo Cameras. *International Conference on Intelligent Robots and Systems*, 2015

[28] A. R. Newcombe et al, KinectFusion: Real-time Dense Surface Mapping and Tracking. *IEEE International Symposium on Mixed and Augmented Reality*, 2011

[29] R. Or-El, G. Rosman, A. Wetzler, R. Kimmel and A. M. Bruckstein, RGBD_Fusion: Real-time High Precision Depth Recovery. *International Conference on Computer Vision and Pattern Recognition*, 2015

[30] O. Woodford and G. Vogiatzis, A Generative Model for Online Depth Fusion. *In proceedings of European Conference of Computer ECCV Vol 7576*, 2012

[31] S. Song, S. Lichtenberg and J. Xiao, SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite, International *Conference on Computer Vision and Pattern Recognition*, 2015

[32] P. Merrel et al. Real-time Visibility-based Fusion of Depth Maps. *IEEE International Conference on Computer Vision* 2007

[33] C. Hane, C. Zach, J. Lim, A. Ranganathan and M. Pollefeys, Stereo Depth Map Fusion for Robot Navigation. *International Conference on Intelligent Robots and Systems*, 2011.

[34] M. Rumpler, A. Wendel and H. Bischof. Probabilistic Range Image Integration for DSM and True-Orthophoto Generation. *Springer Lecture Notes in Computer Science* 2013.

[35] S. Matyunin, D. Vatolin and M. Smirnov, Fast Temporal Filtering of Depth Maps. *International Conference on Computer Graphics, Visualization and Computer vision*, 2011.

[36] C. Unger, E. Wahl, P. Strum and S. Ilic, Probabilistic Disparity Fusion for Real-time Motion Stereo. *Machine Vision and Applications*, Vol 25, 2011.

[37] V. Ntouskos and F. Pirri, Confidence Driven TGV Fusion. *arXiv:1603.09302*, March 2016.

[38] C. Stauffer and W. Grimson, Adaptive Background Mixture Models for Real-time Tracking. *International Conference on Computer vision and Pattern Recognition*, 1999.

[39] A. Geiger, J. Ziegler and C. Stiller, StereoScan: Dense 3D Reconstruction in Real-time. *Intelligent Vehicles Symposium*, 2011.

[40] B. Kitt, A. Geiger and H. Lategahn. Visual Odometry based on Stereo Image Sequences with RANSAC-based Outlier Rejection Scheme. *Intelligent Vehicle Symposium*, 2010.

[41] V. Ntouskos et al. Saliency Prediction in Coherence Theory of Attention. *Biologically Inspired Cognitive Architectures*, 2013

[42] H. Hirschmuller, Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information. *International Conference on Computer Vision and Pattern Recognition*, 2005

[43] A. Geiger, M. Roser and R. Urtasun, Efficient Large Scale Stereo Matching, *Ascian Conference on Computer Vision*, 2010.